References

Bermúdez, J. L. 1997. Defending intentionalist accounts of self-deception. *Behavioral and Brain Sciences* 20: 107–8.

Bermúdez, J. L. 2000. Autoinganno, intenzioni e credenze contraddittorie. Un commento a Mele. *Sistemi Intelligenti* 3: 521–1.

Haight, M. 1980. *A Study of Self-Deception*. Brighton: Harvester Press.

Mele, A. 1997. Real self-deception. *Behavioral and Brain Sciences* 20: 91–102.

Mele, A. 1998. Motivated belief and agency. *Philosophical Psychology* 11: 353–69.

Nozick, R. 1981. *Philosophical Explanations*. Cambridge Mass.: Harvard University Press.

Talbott, W. 1995. Intentional self-deception in a single, coherent self. *Philosophy and Phenomenological Research* 55: 27–74.

Trope, Y. and A. Liberman. 1996. Social hypothesis testing: cognitive and motivational mechanisms. In *Social Psychology: Handbook of Basic Principles*, ed. E. Higgins and A. Kruglanski. 239–70. New York: Guildford Press.

Conservativeness and translation-dependent T-schemes

Jeffrey Ketland

1. The conservativeness of the (restricted) T-scheme (T)

The basic principle of many popular deflationist accounts of truth[1] is the *disquotational T-scheme*

(**T**): ⌜φ⌝ is true in L if and only if φ.

where to avoid semantic paradox we need a restriction: it is natural to insist that the sentence φ must be a sentence of the object or base language L.[2]

In Ketland 1999 I reported the following theorem:

[1] E.g. Paul Horwich's minimalist account of truth (see Horwich 1998). Horwich's theory is based on the *propositional* version of the disquotation scheme, where the singular term on the lhs is a 'that'-clause rather than a quotation (Horwich calls this the 'Equivalence Schema'). Fortunately, it makes no difference to the sort of *formal* investigation I am concerned with: all that matters to the formal investigations is that you have an (injective) mapping from sentences to singular terms. Terms such as 'that snow is white' are singular terms obtained from sentences by an injective mapping.

[2] But see Theorem 4 below, where we introduce (and prove consistent) an *unrestricted* T-scheme, in which the truth-in-L condition for any sentence φ containing 'true-in-L' is some arbitrary absurdity.

Analysis 60.4, October 2000, pp. 319–28.

Theorem 1. The (restricted) T-scheme (**T**) is *conservative* over (almost) any theory Σ in L to which it is added.[3]

Similar arguments yield the conservativeness of the scheme for *satisfaction*:[4]

(**S**): $\forall x_1 \ldots \forall x_n$ ($\ulcorner\varphi\urcorner$ is satisfied in L by the sequence $\langle x \rangle$ iff $\varphi(x_1, \ldots, x_n)$).

In Ketland 1999 (§2) I suggested that the conservativeness of (**T**) is connected to its *deflationary* status (for a similar argument, see Shapiro 1998). Of course, no deflationist has ever put their view exactly like this. But the fact that the (restricted) T-scheme *is* conservative provides an important analysis of a number of features of the truth predicate that the deflationist wants, such as:

(i) The *dispensability* of the truth predicate;
(ii) the *epistemological neutrality* of truth;
(iii) the *contentlessness* of truth.

For example, suppose you have some *non-semantical* theory Σ (perhaps a physical theory) in a language L and you extend it to a theory $\Sigma^+ = \Sigma +$ all L-instances of (**T**). Suppose you deduce in this 'semanticized' theory Σ^+ some *non-semantical* sentence φ (say, a conditional linking initial conditions with final conditions), perhaps using the concept of truth (in-L) in the deduction. Then, the conservativeness theorem above tells you that you can already deduce φ in Σ, *without* invoking the concept of truth. Hence, we have an important sense in which the minimalist truth predicate is *dispensable*. Any non-semantical fact explained *with* (minimalist) truth can be explained *without* it.

Furthermore, the conservativeness of (**T**) implies that (**T**) is consistent with (almost) any internally consistent *non-semantical* theory Σ of the world. Hence, one cannot deduce any *new* non-semantical facts about the world from a deflationist or minimalist theory of truth based on (**T**) alone. For example, imagine some crazy idealist theory Σ which contains a metaphysical axiom or theorem:

[3] *Almost* any theory? Well, Σ must imply the following theory Δ of syntax:

$\Delta = \{\ulcorner\varphi\urcorner \neq \ulcorner\psi\urcorner$: for all L^+ — formulae φ, ψ such that $\varphi \neq \psi\}$

Roughly speaking, 'syntactical identity' of formulae must be strongly represented in Σ. For example, Robinson Arithmetic Q satisfies this constraint, as should any decent theory of syntax for an object language L. Many thanks to Volker Halbach for helping me clear up this point (personal communication).

[4] Strictly, (**S**) below needs to be construed as the *union* of an infinite number of satisfaction schemes $(\mathbf{S})_n$, one scheme for each fixed number n of distinct free variables in the formula $\varphi(x_1, \ldots, x_n)$.

(1) All objects are mental constructions.

Adding the T-scheme (**T**) to this theory Σ won't make any difference to what can or cannot be proved (in the language of Σ) about these objects. I suspect that Tarski would have agreed that this is a precise explication of what he meant when he alluded, in his 1944 paper, to the 'epistemological neutrality' of his work on the semantic conception of truth:

> Thus, we may accept the semantic conception of truth without giving up any epistemological attitude we may have had … The semantic conception is completely neutral toward all these issues. (Tarski (1944) 1999: 140)[5]

The conservativeness of (**T**) implies a kind of contentlessness of minimalist truth. Perhaps that fact accounts for the usual undergraduate response that the scheme (**T**) is 'obvious', even though, when unrestricted, it is actually inconsistent!

I venture that this is *exactly* the sort of (formal) behaviour that the deflationist wants (or should want) from a truth theory. In 1980, Hartry Field argued for a deflationary view of *mathematics* (see Field 1980), driven by the intuitively plausible idea that one shouldn't be able to deduce non-mathematical facts from mathematical theories. By analogy, I suggest that deflationism about *truth* must hinge on the intuitively plausible idea that one shouldn't be able to deduce non-semantical facts (like the fall of communism in 1989) from a theory of truth alone. This analogy is between the properties of mathematical theories crucial to the development of a deflationary view of mathematics (i.e. the (alleged) conservativeness of mathematics) and the properties of certain rather weak truth theories (i.e. the demonstrable conservativeness of minimalist theories of truth).[6]

[5] *If* the idealist/subjectivist/pragmatist wants to add a claim about the *relation* of the concept of truth to some such *epistemic* concept (like knowability, idealized belief, warranted assertability, or provability), then he or she will get into obvious logical trouble with the T-scheme. For example, if you add the axiom,

$\forall x$ (x is true if and only if x is provable in *PA*)

to Peano Arithmetic, *as well* as the (restricted) T-scheme, then the result is *inconsistent*. This is because *PA* implies (correctly) that the gödel sentence *G* is true iff *not* provable in *PA*. Adding an axiom saying *G* is true iff it *is* provable in *PA* leads to an obvious contradiction. In this sense, such theories of truth are ruled out by (**T**).

[6] It was pointed out in Ketland 1999 that Tarski's *full theory of truth* behaves differently (see also Shapiro 1998 for a similar line or argument). To recap briefly, by formalizing Tarski's 1936 axioms for satisfaction, one can construct a rather natural Tarskian truth-theoretic extension of *PA*, which is sometimes called $PA(S)$. (It is important to note that $PA(S)$ expands the induction scheme to include formulae containing the satisfaction predicate. Let $PA(S)_0$ be the system without induction expanded: this is a conservative extension of *PA*.) This axiomatic system contains a

2. *Translational T-scheme*

Why should we only consider disquotational T-sentences? There is no formal difficulty in considering non-disquotational T-sentences, such as

(2) 'Snow is white' is true if and only if snow is *not* white
(3) 'Some cat is on some mat' is true if and only if some cherry is on some tree

or perhaps all instances of non-disquotational schemes, such as,

(4) ⌜φ⌝ is true if and only *not* φ
(5) ⌜φ⌝ is true if and only $0 = 1$

In general, we can proceed by fixing an arbitrary translation mapping f from the object language L to itself, and considering the general meta-linguistic scheme,

$(\mathbf{T})_f$: ⌜φ⌝ is true if and only if f(φ)

Then our disquotational T-scheme (**T**) is just the case of $(\mathbf{T})_f$ when f = I (the identity mapping). The scheme $(\mathbf{T})_f$ is thus a translation-dependent generalization of the usual disquotational T-scheme. And, of course, when the metalanguage (i.e. the language containing the truth predicate; the language *in* which $(\mathbf{T})_f$ is formulated) does not contain the object language, then one must use some kind of translation mapping to form the appropriate ('heterophonic') T-sentences, such as

(6) 'Der Schnee ist weiss' is true-in-German if and only if snow is white

This is the case that arises in the famous application of Tarski's ideas to meaning, translation and interpretation by Donald Davidson (see Davidson 1967). So, $(\mathbf{T})_f$ gives the general form of the non-disquotational T-sentences which the Davidsonian truth theorist is looking for.[7]

---

nice theory of truth for arithmetic formulae – many standard meta-mathematical results can be formalized and proved within *PA*(*S*), that is, the sort of results one learns informal proofs of in an intermediate logic course (e.g. the set of arithmetic truths is not recursively axiomatizable). The important thing is that *PA*(*S*) is a *non-conservative* extension of *PA* (indeed, you can *deduce* the gödel sentence *G* for *PA* in *PA*(*S*)). There is (as Shapiro and I went on to spell out) a very deep sense in which the *minimalist* truth theory (Horwich-style) is deeply *inadequate* as a representation of our grasp of the concept of truth, as compared with a system like *PA*(*S*). Thus, when the technical dust settles, we have the following philosophical conditional: *if* deflationism turns on conservativeness, and Tarski's theory is (in certain circumstances) non-conservative, *then* deflationism is wrong.

[7] Except that Davidson works 'backwards'. One starts with the Tarski-style truth theory that implies a system of heterophonic T-sentences, and then one attempts to 'extract' the relevant translation mapping f that generated these sentences.

Here, I want to discuss the simpler case, more usual in logical studies, where the metalanguage is an extension of the object language L. The usual (or even 'natural') metalanguage L^+ is the object language L plus a new predicate *Tr(x)*, governed at the very least by the disquotation scheme (**T**).[8]

3. *Conservativeness of the restricted translational T-scheme*

Having explained the philosophical significance of the notion of conservativeness, I want to report an interesting fact. This translation-dependent scheme $(\mathbf{T})_f$ is also conservative over (almost) any theory Σ, *no matter how insane the translation mapping f is*!

The theorems below are somewhat technical and I shall briefly go through the details. L is a language (the base language) which contains the usual first-order language of arithmetic with signature $\underline{0}$, s, $+$, $\times$. Extend L to a language L^+ by introducing a new monadic predicate symbol *Tr*. Let the map #: $L^+ \rightarrow \mathbf{N}$ be some gödel coding function, where $\mathbf{N}$ is the set of natural numbers. Let *Sent*(L) be the set of L-sentences and let *Form*(L) be the set of L-formulae (and correspondingly for L^+).

Let Σ be a theory in L which proves each 'distinctness' formula $\ulcorner\varphi\urcorner \neq \ulcorner\psi\urcorner$, for each pair of distinct L^+-formulae φ and ψ, where $\ulcorner\varphi\urcorner$ is the usual gödel *numeral* for φ relative to the coding function # (i.e. $\ulcorner\varphi\urcorner$ is the term $s \ldots s(\underline{0})$, with $\#\varphi$ iterations of s). If this theory Σ has a model M at all, then all the *syntax* of L^+ can be 'coded' as elements of dom(M). (If M satisfies certain other axioms as well, it is sometimes called an 'acceptable structure'). Let us describe how L^+-formulae (and terms) are coded within this model M. The fact that $M \models \ulcorner\varphi\urcorner \neq \ulcorner\psi\urcorner$ for all distinct formulae φ, ψ implies that dom(M) must be infinite. Since $\vdash \mathrm{dom}(M) \vdash \leq \aleph_0$, one can then define a new (injective) coding function #*: *Form*(L^+) $\rightarrow$ dom(M) which maps each L^+-formula φ into dom(M) such that $(\ulcorner\varphi\urcorner)^M = \#^*(\varphi)$. To achieve this, we define $\#^*(\varphi)$ to be $g^n(\underline{0}^M)$ (g applied n times), where the function g is just $(s)^M$ (the interpretation of the successor symbol in M) and n is the usual code, $\#\varphi$. The fact that $M \models \ulcorner\varphi\urcorner \neq \ulcorner\psi\urcorner$ for all distinct φ, ψ, implies that this coding function #* is injective. Hence, each L^+-formula φ can be coded as a distinct element $\#^*(\varphi) \in \mathrm{dom}(M)$.

[8] When the base theory Σ in L is sufficiently rich (e.g. Peano arithmetic), this scheme cannot be *interpreted* within the base language, by Tarski's Indefinability Theorem: there is no (arithmetic) L-formula $\Psi(x)$ such that $PA \vdash \Psi(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$, for all L-sentences φ. This sort of fact already points to the *irreducibility* of (even minimalist) truth. In general, if X is a sufficiently rich domain of truths, then the *concept* of truth in X is not definable in X. For example, on plausible assumptions, physical truth is not definable in physics. Set-theoretical truth is not definable in set theory. And so on. This implies that, despite the claims of semantic naturalists, truth (and associated alethic notions like reference and satisfaction) is conceptually *irreducible*.

Theorem 2. Let f: $Sent(L) \rightarrow Sent(L)$ be an arbitrary map. Consider the restricted scheme

$$(\mathbf{T})_f\text{: } Tr(\ulcorner \varphi \urcorner) \leftrightarrow f(\varphi)$$

where φ is any L-sentence. Then, $\Sigma \cup (\mathbf{T})_f$ is a conservative extension of Σ.

Proof. If Σ is inconsistent, then any extension is conservative. So, assume that Σ is consistent and pick any model M of Σ. (We use the model-theoretic method of proving conservativeness.[9]) We define an expansion M^+ as follows.[10] Let $M^+[Tr]$ be the set $\{\#^*(\varphi) \in \text{dom}(M)\text{: } \varphi \in Sent(L)$ and $M \models f(\varphi)\}$. Then, clearly, $M^+ \models (\mathbf{T})_f$. Thus, the scheme holds in M^+. It is a well-known result of mathematical logic that if any model M of a theory Σ can be *expanded* to a model M^+ of some extension Σ^+, then Σ^+ must be a conservative extension of Σ.[11] Hence, $\Sigma \cup (\mathbf{T})_f$ is a conservative extension of Σ. □

[9] A *proof-theoretic* method is also available (and in some cases preferable). The method is to show how to convert any proof Γ of an L-sentence φ in the extension $\Sigma \cup (\mathbf{T})_f$ to a new proof Γ^* in Σ. This proceeds by finding a suitable L-predicate $\Psi(x)$ which behaves like a restricted truth predicate and using this to reinterpret each instance of the T-scheme that occurs in the original proof Γ as an L-formula which is provable in Σ. In brief, if the T-instances that occur in the original proof Γ are $Tr(\ulcorner \varphi_i \urcorner) \leftrightarrow f(\varphi_i)$, for some finite set $\{\varphi_1, \ldots, \varphi_n\}$ of L-sentences, then we choose $\Psi(x)$ to be the L-formula

$$(x = \ulcorner \varphi_i \urcorner f(\varphi_i)) \vee \ldots \vee (x = \ulcorner \varphi_n \urcorner \wedge f(\varphi_n)).$$

Then one can show that each formula $\Psi(\ulcorner \varphi_i \urcorner) \leftrightarrow f(\varphi_i)$ is a theorem of Σ. Replacing each T-instance $Tr(\ulcorner \varphi_i \urcorner) \leftrightarrow f(\varphi_i)$ in the original proof Γ by $\Psi(\ulcorner \varphi_i \urcorner) \leftrightarrow f(\varphi_i)$ yields a new proof Γ^* which is a proof of φ in Σ. (More exactly, the sequence Γ^* of L-formulae can be enlarged to a genuine proof of φ.)

[10] An 'expansion' of a model corresponds to an *extension* of a language L. Roughly, a model M is an ordered tuple $(D, R_1, \ldots, R_n)$. An *expansion* of M is obtained by 'adding' a new *relation* R^* to obtain a richer structure $(D, R_1, \ldots, R_n, R^*)$ and correspondingly a new primitive relation symbol to the language associated with M. If the new relation R^* can already be defined in M (by some L-formula $\varphi(x)$), we say that the expansion is a *definitional* expansion. For example, adding the usual order relation $<$ to the natural number structure $(\mathbf{N}, 0, s, +, \times)$ is a definitional expansion.

[11] To see this, suppose that every model M of Σ can be expanded to a model M^+ of Σ^+ and suppose that not-$(\Sigma \models \varphi)$, where φ is an L-sentence. By the completeness theorem, there is a model M of Σ in which φ is false. By assumption, M can be expanded to a model M^+ of Σ^+ where φ is also false. Hence, not-$(\Sigma^+ \models \varphi)$. (The converse of the result is not true: Σ^+ can be a conservative extension of Σ even though there are models of Σ which cannot be expanded to a model of Σ^+.) For example, see Hodges 1997: 58–59. In general, the level of mathematical logic assumed in this article is standard intermediate (classical) first-order logic, e.g. Boolos and Jeffrey 1989.

One can obtain a similar result for a satisfaction predicate (or what Putnam 1981 calls a reference predicate). One can (partially) define a 'satisfaction' predicate $Sat(x, y)$ by an f-dependent satisfaction scheme like

$$(\mathbf{S})_f\colon \forall x_1 \ldots \forall x_n\, [Sat(\ulcorner \varphi \urcorner, \langle x \rangle) \leftrightarrow f(\varphi(x_1, \ldots, x_n)].$$

It is not hard to show that the result of adding the 'satisfaction' scheme $(\mathbf{S})_f$ to Peano Arithmetic *PA* (and thus any extension) is also conservative, by a construction analogous to the one above. In brief, let M be a model of *PA* (this M needn't be standard) and let $S_f(M)$ be the f-dependent 'satisfaction' relation on the structure M. That is, the class of pairs $(\#\varphi, m)$ where φ is an L-formula, the number m codes a finite sequence $\langle s \rangle$ of elements of M, and $f(\varphi)$ is satisfied in M by that sequence $\langle s \rangle$. Then you define the natural expansion M^+ by setting the extension $M^+[Sat]$ to be the relation $S_f(M)$. This expansion M^+ satisfies the f-dependent 'satisfaction' scheme.

To illustrate, consider adding to *PA* all the sentences

$$Tr(\ulcorner \varphi \urcorner) \leftrightarrow \varphi^*$$

where φ doesn't contain $Tr(x)$, and φ^* is obtained from φ by swapping any occurrences of the symbols + and ×. The result is a conservative extension.

For a non-mathematical example, consider adding to non-semantical English all the sentences,

(7) 'Snow is white' is true if and only if snow is not white
(8) 'Grass is green' is true if and only if grass is not green etc.

The extension of any non-semantical theory in English obtained by adding these axioms is conservative.[12]

4. *Conservativeness of unrestricted translational T-schemes*

We have shown that the restricted translational T-scheme $(\mathbf{T})_f$ is conservative for *any* translation mapping f. What about the *unrestricted* translational T-scheme? This is the scheme,

$$(\mathbf{T})_f{}^*\colon Tr(\ulcorner \varphi \urcorner) \leftrightarrow f(\varphi)$$

[12] I hope it is obvious that the quasi-semantical concepts being introduced by these 'twisted' schemes are *not* truth or satisfaction. In short, I would maintain there is really something *special* about the *disquotation* scheme(s), where $f = \mathrm{I}$. One way to see this is to consider, instead of a single truth predicate $Tr(x)$, a distinct predicate $Tr_f(x)$ for each translation mapping f. Then the scheme $(\mathbf{T})_f$ is

$$(\mathrm{T})_f\colon Tr_f(\ulcorner \varphi \urcorner) \leftrightarrow f(\varphi).$$

The claim would be that there is something special about the predicate $Tr_{\mathrm{I}}(x)$ (the case where $f = \mathrm{I}$). I lack the space to discuss this issue here, but I hope to discuss it more fully at a later date.

where φ can now be *any* L^+-sentence, even one containing the predicate $Tr(x)$. More exactly, the appropriate set of instances of the scheme $(\mathbf{T})_f^*$ is the set $\{Tr(\ulcorner\varphi\urcorner) \leftrightarrow f(\varphi)\colon \varphi \in Sent(L^+)\}$. Suppose we set $f = \mathrm{I}$. Then it is a well-known consequence of the Diagonalization Lemma that $(\mathbf{T})_\mathrm{I}^*$ is *inconsistent* when added to theories like *PA* (or decent theories of syntax). Consider the formal theory $PA \cup (\mathbf{T})_\mathrm{I}^*$. By the Diagonalization Lemma there must exist a fixed point 'liar' formula λ such that $PA \cup (\mathbf{T})_\mathrm{I}^* \vdash \neg Tr(\ulcorner\lambda\urcorner) \leftrightarrow \lambda$. But we already have that $PA \cup (\mathbf{T})_\mathrm{I}^* \vdash \neg Tr(\ulcorner\lambda\urcorner) \leftrightarrow \lambda$, so $PA \cup (\mathbf{T})_\mathrm{I}^*$ must be inconsistent. Thus, the choice $f = \mathrm{I}$ (which is roughly equivalent to the *naive disquotational conception of truth*) is ruled out by logical considerations. At the very least, this implies (at least if we adhere to classical logic) that the *full* disquotation scheme must be restricted in some way.

However, there is an interesting intermediate result. For a *large class of choices* of translation mapping *f*, we can prove that the associated *unrestricted* T-scheme is conservative (and thus consistent) also.

Theorem 3. Same assumptions as in Theorem 2. Let $f\colon Sent(L^+) \to Sent(L)$ be any translation mapping (N.B. the range of *f* is the set of *L*-sentences). Consider the unrestricted translational scheme

$$(\mathbf{T})_f^*\colon Tr(\ulcorner\varphi\urcorner) \leftrightarrow f(\varphi)$$

where φ is now any L^+-sentence. Then, $\Sigma \cup (\mathbf{T})_f^*$ is a conservative extension of Σ.

Proof. As before, choose a model *M* of Σ and fix an injective coding function $\#^*\colon Sent(L^+) \to \mathrm{dom}(M)$. Define an expansion M^+ similar to the previous one: let $M^+[Tr]$ be the set $\{\#^*(\varphi) \in \mathrm{dom}(M)\colon \varphi \in Sent(L^+)$ and $M \models f(\varphi)\}$. (Notice that this makes good sense because $f(\varphi)$ is always an *L*-sentence, even when φ is an L^+-sentence.) Again, we can show that $M^+ \models (\mathbf{T})_f^*$. If $\varphi \in Sent(L)$ then proceed as before. If $\varphi \in \mathrm{Sent}(L^+/L)$, then $M^+ \models Tr(\ulcorner\varphi\urcorner)$ iff $\#^*(\varphi) \in M^+[Tr]$ iff $M \models f(\varphi)$. Because $f(\varphi)$ is an *L*-sentence, we have $M \models f(\varphi)$ iff $M^+ \models f(\varphi)$. Hence, $M^+ \models Tr(\ulcorner\varphi\urcorner) \leftrightarrow f(\varphi)$. Hence, $\Sigma \cup (\mathbf{T})_f^*$ is a conservative extension of Σ. □

Again, the same trick works for a satisfaction predicate. What is crucial in this theorem is that the *range* of the translation function *f* be the class of *L*-sentences. For example, we can translate the L^+-predicate $Tr(x)$ as the formula $x = \underline{0}$. In general, we can translate the 'truth' predicate $Tr(x)$ as any formula of *L* (with just *x* free). Theorem 2 is a consequence of Theorem 3. An application of Theorem 3 is this:

Theorem 4. Same assumptions as in Theorem 2. Define a translation mapping *f* by:

(i) $f(\varphi) = \varphi$, for any $\varphi \in Sent(L)$;
(ii) $f(\varphi) = (\underline{0} \neq \underline{0})$, for any $\varphi \in Sent(L^+/L)$.

Then, $\Sigma \cup (\mathbf{T})_f{}^*$ is a conservative extension of Σ.

Proof. The range of the mapping f on $Sent(L^+)$ is $Sent(L)$. Apply Theorem 3. □

There is a rough sense in which this choice of f is Tarski's choice. For this mapping f is disquotational on L-sentences, but maps every L^+-sentence to an absurdity in L. Intuitively, the idea here is that the Tarskian truth predicate $Tr(x)$ really means 'x is true in L', where L is the language not containing the truth predicate (i.e. L is the base or object language). Clearly, if a formula φ *contains* the truth predicate, then it is *not a sentence of* L (since the truth predicate is not definable in L) and thus is *not true* in L either.[13] Thus, no such φ is true in L and thus the correct truth-in-L condition for φ is the absurdity $\underline{0} \neq \underline{0}$ (or $\perp$, if you like). That is what the translation mapping f in Theorem 4 achieves.[14]

University of Nottingham
University Park, Nottingham NG7 2RD, UK
Jeffrey.Ketland@nottingham.ac.uk

References

Boolos, G. and R. Jeffrey. 1989. *Computability and Logic*. 3rd ed. Cambridge: Cambridge University Press.

Davidson, D. 1967. Truth and meaning. *Synthese* 17: 304–23. Reprinted in D. Davidson 1984, *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press.

Field, H. 1980. *Science Without Numbers*. Princeton: Princeton University Press.

Hodges, W. 1997. *A Shorter Model Theory*. Cambridge: Cambridge University Press.

Ketland, J. 1999. Deflationism and Tarski's paradise. *Mind* 108: 69–94.

Ketland, J. 2000. A proof of the (strengthened) Liar formula in a semantical extension of Peano arithmetic. *Analysis* 60: 1–4.

Horwich, P. 1998. *Truth*. 2nd ed. Oxford: Oxford University Press.

Putnam, H. 1981. *Reason, Truth and History*. Cambridge: Cambridge University Press.

Shapiro, S. 1998. Proof and truth – through thick and thin. *Journal of Philosophy* 95: 493–522.

Tarski, A. 1936. Der Wahrheitsbegriff in den formalisierten Sprachen. *Studia Philosophica* 1: 261–405. English translation by J. H. Woodger, The concept of truth in

[13] Such an L^+-formula φ *containing* the truth(-in-L) predicate $Tr_L(x)$ might easily be true in the metalanguage L^+, even though is it not true in the object/base language L. Indeed, this is exactly what happens with the strengthened liar formula λ (i.e. the 'fixed point formula' such that $\lambda \leftrightarrow \neg Tr_L(\ulcorner \lambda \urcorner)$ is a theorem). This example is worked out in some detail in Ketland 2000, where it is shown that the liar formula λ is not only true (in L^+), but is in fact *provable* in the semantical metatheory $PA(S)$.

[14] I would like to thank Volker Halbach, Vann McGee, Michael Clark and an anonymous referee for comments on the ideas in this article.

formalized languages, appeared in A. Tarski 1956, *Logic, Semantics and Metamathematics: Papers by Alfred Tarski 1923–1938*. Oxford: Clarendon Press.

Tarski, A. 1944. The semantic conception of truth and the foundations of semantics. *Philosophy and Phenomenological Research* 4: 342–60. Reprinted in *Truth*, ed. S. Blackburn and K. Simmons. 1999. Oxford: Oxford University Press.

On Soames's solution to the sorites paradox

TERESA ROBERTSON

Scott Soames (1999, ch. 7) has recently offered a new solution to the sorites paradox. Although this solution has some appeal, it seems to me that, short of some substantial revision, it fails.

1. Presentation of Soames's solution

Soames's solution to the sorites paradox turns on two features of vague predicates: (i) they are (at least potentially) partially defined and (ii) they are context-sensitive. A partially defined predicate, 'is *F*' say, is one whose extension and antiextension are mutually exclusive but not jointly exhaustive. For any object *o* that is in neither the extension nor the antiextension, both the claim that *o* is *F* and the claim that *o* is not *F* should be rejected. There is (at least potentially) a truth-value gap.[1] Given the understanding of the material conditional that is provided by the strong Kleene tables,[2] this means that for any standard sorites paradox of the form of the (implausibly short) one displayed in §2, the conditionals near the beginning will be true; the ones in the middle will lack truth values – that is, will be 'undefined'; and the ones toward the end will be true, since their antecedents will be false. None will be outright false. So the conclusions of sorites arguments can be avoided by rejecting some of the conditional premisses without thereby being forced to accept their negations.

By itself, this is not very satisfactory. Soames considers two objections.

Objection 1. If partial definition were the whole story and *o* were an object in the undefined range of the predicate, then we would have no

[1] The gap would be merely potential in the case of 'is bald' if, for example, all people had no hair.

[2] Where '*U*' is used for neither *T* nor *F*, we have the following: (*T*, *T*) yields *T*; (*T*, *U*) yields *U*; (*T*, *F*) yields *F*; (*U*, *T*) yields *T*; (*U*, *U*) yields *U*; (*U*, *F*) yields *U*; (*F*, *T*) yields *T*; (*F*, *U*) yields *T*; (*F*, *F*) yields *T*.

Analysis 60.4, October 2000, pp. 328–34.